3 COURSE

Sampling and Estimation

Sampling and estimation are fundamental in inferential statistics. They allow us to draw conclusions about a large population based on the analysis of a smaller representative subset (sample). In this chapter, we introduce the essential notions of sampling, random sampling distributions, and estimation of unknown parameters.

1 Sampling

3.1.1 Concept of Sampling

Definition 1.1

Consider a population Ω of size N. A **sample** is a subset of this population. A sample of size n is thus a list of n individuals $(\omega_1, \omega_2, ..., \omega_n)$ drawn from the parent population.

Example 1.1

Consider a population composed of 5 students. We are interested in the weekly time devoted by each student to studying statistics.

$$\Omega = \{A, B, C, D, E\}, \quad N = 5$$

Student	Study Time (h)
A	7
B	3
C	6
D	10
E	4

Definition 1.2

Sampling is the process of selecting samples. The ratio t of the sample size n to the population size N from which it is drawn is called the **sampling rate** or **sampling fraction**, i.e.

$$t = \frac{n}{N}$$

Example 1.2

If we draw samples of size 2, then $t = \frac{2}{5}$ (see Example 1.1).

Definition 1.3

A **random sample** is a selection of n individuals from a parent population such that all possible combinations of n individuals have the same probability of being selected. Other types of sampling exist, but we will focus exclusively on random sampling.

Remark

We aim to describe a qualitative or quantitative characteristic C of a population Ω by studying the results obtained from a sample of size n.

Example 1.3

1. For a given population, we may study quantitative characteristics such as weight or height.
2. For a given population, we may study qualitative characteristics such as eye color or hair color.
3. In the initial example, the characteristic studied is the weekly time devoted to studying statistics.

Definition 1.4

Let C be a quantitative characteristic defined on a parent population Ω. C is the realization of a random variable X defined on Ω:

$$X : \Omega \to \mathbb{R}, \quad \omega_i \mapsto X(\omega_i) = x_i$$

A **sample of values** of X is the list of observed values $(x_1, x_2, ..., x_n)$ taken by X on a sample $(\omega_1, ..., \omega_n)$ of the population Ω. The coordinates can be regarded as realizations of a random vector $(X_1, ..., X_n)$ called an n-sample of X, where the X_i are independent and identically distributed (i.i.d.) random variables with the same distribution as X.

Definition 1.5

Any random variable that can be expressed in terms of the random variables $X_1, ..., X_n$ is called a **statistic**.

Example 1.4

X_i and $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ are examples of statistics.

Remark

If we extract several samples of the same size n, the results we obtain will vary depending on the sample considered. We call this variability **sampling fluctuations**. To make reliable inferences about the parent population, we must study the probability laws governing these fluctuations.

3.1.2 Sampling Distributions

Sample Mean and Sample Variance

Definition 1.6

Consider a population Ω whose elements possess a quantitative characteristic C that is the realization of a random variable X with expectation μ and standard deviation σ. Assume the population is infinite or that sampling is done with replacement.
We draw a sample $(X_1, ..., X_n)$ from X, giving observed values $(x_1, ..., x_n)$. The **sample mean** is given by:

$$\overline{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

The corresponding random variable is:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Similarly, the **sample variance** is:

$$v = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

and the associated random variable:

$$V = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

We define the random variable S^2, called the **unbiased sample variance**, as:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{n}{n-1}V$$

3.1.3 Sample Proportion

Definition 1.7

Sometimes, the characteristic to be estimated is not quantitative but qualitative. In this case, we seek the proportion p of individuals in the population possessing that characteristic. The proportion p is estimated from the results obtained in a sample of size n.
The observed proportion f in a sample is the realization of a random variable F, representing the frequency of appearance of this characteristic in the sample. F is called the **sample proportion** or **statistical frequency**:

$$F = \frac{K}{n}$$

where K is the random variable counting the number of occurrences of the characteristic in the sample of size n. By definition, $K \sim B(n, p)$, so that:

$$E(K) = np, \quad Var(K) = npq \quad \text{with } q = 1 - p.$$

Therefore,

$$E(F) = p, \quad Var(F) = \frac{pq}{n}.$$

Remark

For $n \geq 30$, with $np \geq 15$ and $nq \geq 15$, F can be approximated by a normal distribution:

$$F \sim N\left(p, \sqrt{\frac{pq}{n}}\right).$$